

### Exercise 1: L0 Regularization

Consider the regression learning setting, i.e.,  $\mathcal{Y} = \mathbb{R}$ , and feature space  $\mathcal{X} = \mathbb{R}^p$ . Let the hypothesis space be the linear models:

$$\mathcal{H} = \{f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} \mid \boldsymbol{\theta} \in \mathbb{R}^p\}.$$

Suppose your loss function of interest is the L2 loss  $L(y, f(\mathbf{x})) = \frac{1}{2}(y - f(\mathbf{x}))^2$ . Consider the  $L_0$ -regularized empirical risk of a model  $f(\mathbf{x} \mid \boldsymbol{\theta})$ :

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 = \frac{1}{2} \sum_{i=1}^n \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}\right)^2 + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0}.$$

Assume that  $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ , which holds if  $\mathbf{X}$  has orthonormal columns. Show that the minimizer  $\hat{\boldsymbol{\theta}}_{L_0} = (\hat{\theta}_{L_0,1}, \dots, \hat{\theta}_{L_0,p})^\top$  is given by

$$\hat{\theta}_{L_0,i} = \hat{\theta}_i \mathbb{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}, \quad i = 1, \dots, p,$$

where  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^\top = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  is the minimizer of the unregularized empirical risk (w.r.t. the L2 loss).

For this purpose, use the following steps:

(i) Derive that

$$\arg \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^p -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda \mathbb{1}_{|\theta_i| \neq 0}.$$

(ii) Note that the minimization problem on the right-hand side of (i) can be written as  $\sum_{i=1}^p g_i(\theta_i)$ , where

$$g_i(\theta) = -\hat{\theta}_i \theta + \frac{\theta^2}{2} + \lambda \mathbb{1}_{|\theta| \neq 0}.$$

What is the advantage of this representation if we seek to find the  $\boldsymbol{\theta}$  with entries  $\theta_1, \dots, \theta_p$  minimizing  $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ ?

(iii) Consider first the case that  $|\hat{\theta}_i| > \sqrt{2\lambda}$  and infer that for the minimizer  $\theta_i^*$  of  $g_i$  it must hold that  $\theta_i^* = \hat{\theta}_i$ .

*Hint:* Show that  $g_i(\hat{\theta}_i) < 0 = g_i(0)$  and argue that the minimizer must have the same sign as  $\hat{\theta}_i$ .

(iv) Derive that  $\theta_i^* = \hat{\theta}_i \mathbb{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}$ , by using (iii) (and also still considering the case  $|\hat{\theta}_i| > \sqrt{2\lambda}$ ).

(v) Consider the complementary case of (iii) and (iv), i.e.,  $|\hat{\theta}_i| \leq \sqrt{2\lambda}$ , and infer that for the minimizer  $\theta_i^*$  of  $g_i$  it must hold that  $\theta_i^* = 0$ .

*Hint:* What is  $g_i(0)$ ? Consider  $\tilde{g}_i(\theta) = -\hat{\theta}_i \theta + \frac{\theta^2}{2} + \lambda$  which is the smooth extension of  $g_i$ . What is the relationship between the minimizer of  $g_i$  and the minimizer of  $\tilde{g}_i$ ?

### Exercise 2: Regularization

(a) Simulate a data set with  $n = 100$  observations based on the relationship  $Y = \sin(x_1) + \varepsilon$  with noise term  $\varepsilon$  following some distribution. Simulate  $p = 100$  additional covariates  $x_2, \dots, x_{101}$  that are not related to  $Y$ .

(b) On this data set, use different models (and software packages) of your choice to demonstrate

- overfitting and underfitting;

- $L1$ ,  $L2$  and elastic net regularization;
- the underdetermined problem;
- the bias-variance trade-off;
- early stopping using a simple neural network.